

RNA Editing Analysis

Thomas Wieland

Helmholtz Zentrum München
Institute of Human Genetics

RNA editing - Introduction

- Most prevalent type: Adenosine to Inosine (A->I) editing by ADAR enzymes
 - ADAR acts on dsRNA
 - Most known sites are in UTRs and intronic retrotransposon elements, especially Alu elements
 - I is read as G → we're looking for A->G variants

Further information:

[Farajollahi and Maas, 2010](#) and [Nishikura, 2010](#)

RNA editing – Howto discover

Find differences between DNA-seq and RNA-seq,
but it's not that simple...

RESEARCH ARTICLES

Genomes Project (19) sequenced the DNA of these individuals. Comparison of sequence data from these two projects showed high concordance.

Widespread RNA and DNA Sequence Differences in the Human Transcriptome

Mingyao Li,^{1*} Isabel X. Wang,^{2*} Yun Li,^{3,4} Alan Bruzel,² Allison L. Richards, Jonathan M. Toung,⁶ Vivian G. Cheung^{2,7,8†}

The transmission of information from DNA to RNA is a critical process. We compared human B cells of 27 individuals to the corresponding DNA sequences from the same individuals and uncovered more than 10,000 exonic sites where the RNA sequences do not match the DNA. 12 possible categories of discordances were observed. These differences were non-synonymous and were found in multiple individuals and in different cell types, including peripheral blood mononuclear cells. Using mass spectrometry, we detected peptides that are translated from the RNA sequences and thus do not correspond exactly to the DNA sequences. These findings indicate that differences in the human transcriptome provide a yet unexplored aspect of genome

OPEN ACCESS Freely available online



Very Few RNA and DNA Sequence Differences in the Human Transcriptome

Daniel R. Schrider^{1,2*}, Jean-Francois Gout^{1*}, Matthew W. Hahn^{1,2}

¹ Department of Biology, Indiana University, Bloomington, Indiana, United States of America, ² School of Informatics and Computing, Indiana University, Bloomington, Indiana, United States of America

Abstract

RNA editing is an important cellular process by which the nucleotides in a mature RNA transcript are altered to cause them to differ from the corresponding DNA sequence. While this process yields essential transcripts in humans and other organisms, it is believed to occur at a relatively small number of loci. The rarity of RNA editing has been challenged by a recent comparison of human RNA and DNA sequence data from 27 individuals, which revealed that over 10,000 human exonic sites appear to exhibit RNA-DNA differences (RDDs). Many of these differences could not have been caused by either of the two previously known human RNA editing mechanisms—ADAR-mediated A→G substitutions or APOBEC1-mediated C→U switches—suggesting that a previously unknown mechanism of RNA editing may be active in humans. Here, we reanalyze these data and demonstrate that genomic sequences exist in these same individuals or in the human genome that match the majority of RDDs. Our results suggest that the majority of these RDD events were observed due to accurate transcription of sequences paralogous to the apparently edited gene but differing at the edited site. In light of our results it seems prudent to conclude that if indeed an unknown mechanism is causing RDD events in humans, such events occur at a much lower frequency than originally proposed.

RNA editing – Variant calling

- Whole-genome multisample calling is still running
→ we did additional multisample calling at ~42,000 known RNA editing sites obtained from [DARNED](#)

```
samtools mpileup -C0 -m3 -F0.0002 -E -l chr1.darned.bed -d999999 \  
-q20 -DSuf hg19.fa -b inputBams | \  
bcftools view -cgv -l chr1.darned.bed - 2> chr1mpileup.log | \  
vcfutils.pl varFilter -Q25 -d3 -D4999500 -a2 -w10 -W10 -10.0001 \  
-21e-400 -30 -40.0001 -p > chr1not_filtered.vcf 2> chr1filtered.vcf
```

→ 24,680 variants called

RNA editing – Filters

- Passed the varFilter script
- Minimum median coverage of 10
- At least 10 non-reference genotypes
- Minimum variant quality of 100
- No genomic variants
 - No corresponding variant in genotype files
 - Lies in 1000G Project „accessible region“

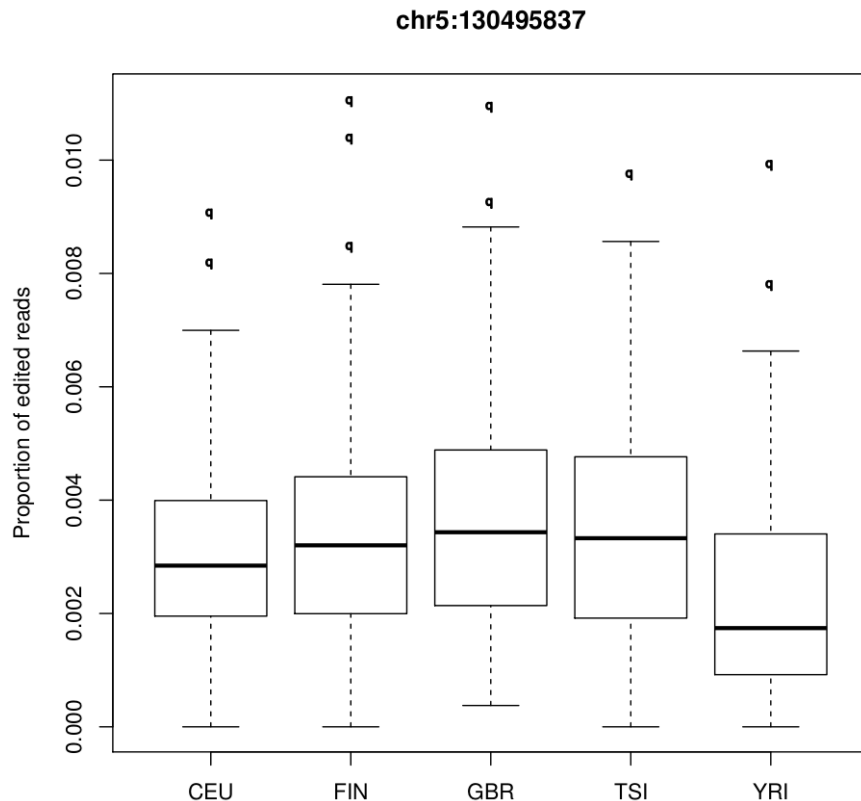
→ **113** variants remain

RNA editing – Results

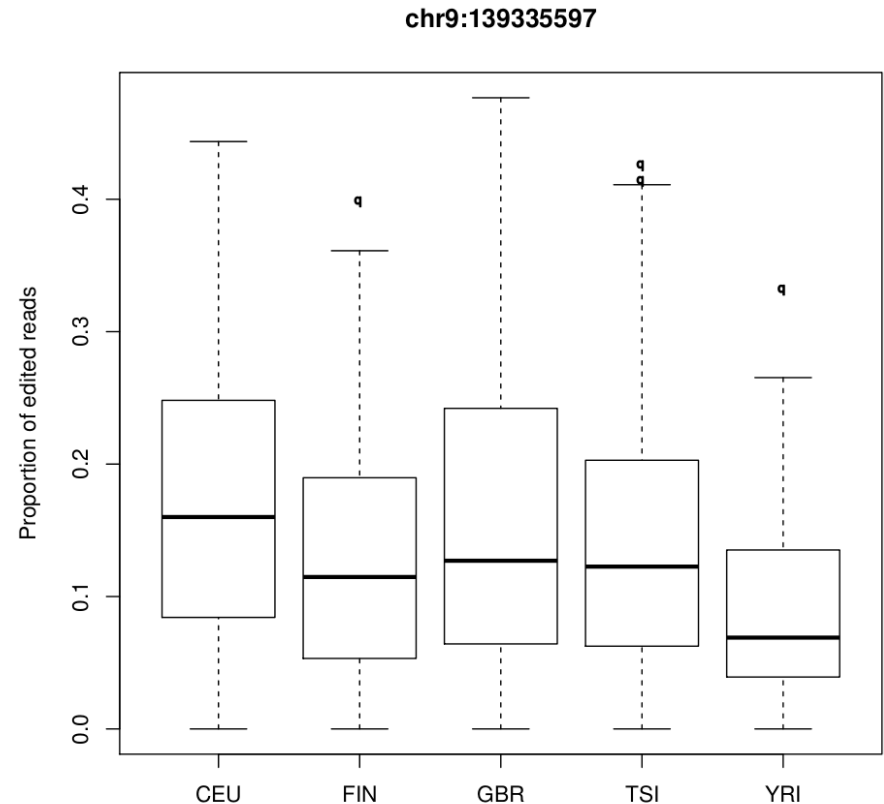
- VCF file with resulting variants as produced by SAMtools with three additional fields per sample:
 - GGT: genotype from genotype files (should be empty)
 - GGL: genotype likelihood from genotype files (should be empty)
 - DP4: reference/non-reference reads per sample
- Stats file with number of edited samples per site
- Proportion file with proportion of edited reads per site per sample

RNA editing – Population specific editing

11 sites are significantly different (ANOVA), e.g.



anova p-value = 2.108099081415e-06



anova p-value = 2.88151011712264e-07